УДК 577.21

# НАРУШЕНИЕ ВТОРОГО ПРАВИЛА ЧАРГАФФА У МИТОХОНДРИАЛЬНЫХ ГЕНОМОВ И ЕГО СВЯЗЬ С ТАКСОНОМИЕЙ НОСИТЕЛЯ

**Зайцева Н.А.**
**Научный руководитель — д-р. физ.-мат. наук Садовский М.Г.**

***Сибирский федеральный университет***

## 1 Введение

Центральной задачей системной биологии является выявление связи между структурой нуклеотидных последовательностей и функцией, которую они кодируют, либо таксономией носителей этих последовательностей. Настоящая работа посвящена выявлению этой связи. Данная задача может быть решена различными способами. В рамках настоящей работы мы будем изучать корреляцию( связи) между группами( классами), выделяемыми исключительно статистическим свойствам, с классами, определяемыми посредством тех или иных биологически обусловленных связей между объектами.

Под структурой в рамках настоящей работы будет пониматься словарь толщиной 3 (т. е., частоты триплетов, наблюдаемые в последовательности). Введём понятие частоты. Для этого рассмотрим непрерывную последовательность символов, состоящую только из четырёхбуквенного алфавита {A, C, G, T} длины $N$. Любую строку $\omega = \nu_1\nu_2\nu_3\ldots\nu_q$ назовем словом длинной $q$. Предметом исследований представленной работы являются слова для которых $q = 3$. Подсчитаем число копий $n_\omega$ каждого слова длины 3. Таким образом, определим частоту как отношение числа копий данного слова к числу всех триплетов

$$f_\omega = \frac{n_\omega}{N} .$$

Исследование связи структуры и таксономии проводилось на генетическом материале митохондрий. Эти геномы весьма малы- характерная длина составляет порядка $10^5$ нуклеотидов. Их особенности заключаются в отсутствии кроссинговера и высокой скорости мутирования, что позволяет отследить число предков на большое количество поколений и изучить филогению (эволюционное родство) живых организмов. Кроме того в геномах митохондрий наблюдается самый высокий уровень нарушений второго правила Чаргаффа. Поэтому, изучение нарушения симметрии, а также влияние этих нарушений на связь между структурой генома и видом его носителя, является одной из задач, представленной работы.

## 2 Материалы и методы

Геномы были взяты из EMBL-банка (www.ebi.ac.uk/genoms/organelle.html), их общее число превышает 2000. Каждый геном был представлен в виде точки в 63-мерном либо 32-мерном пространстве. Для построения распределения была использована программа ViDaExpert (http://bioinfo-out.curie.fr/projects/vidaexpert/), проводящая классификацию методом динамических ядер и методом упругих карт. Ключевой задачей работы является проведение сравнительного анализа состава классов, полученных с помощью применения данных типов классификации.

**3 Результаты и обсуждение**

Основным механизмом исследования связи между таксономическим положением носителя генома и его структурой является построение автоматических классификаций для разных пространств. В настоящей работе мы использовали четыре пространства:

- 63-мерное пространство частот. Один из триплетов был исключён, поскольку на частоты триплетов наложена линейная связь. Выбор конкретного исключённого триплета особенного значения не имеет.
- 32-мерное подпространство частот, включающее в себя те триплеты, которые являются «левыми» в комплиментарных палиндромических парах;
- 32-мерное пространство частот, включающее в себя те триплеты, которые являются «правыми» в комплиментарных палиндромических парах, и, наконец
- -32-мерное пространство разностей частот двух триплетов, составляющих комплиментарные палиндромы.

**3.1 Исследование состава класса в 63- мерном пространстве**

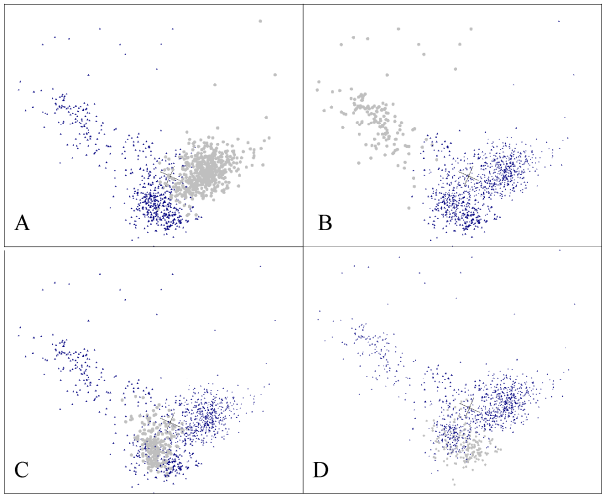


Рис. 2:Распределение 1132 геномов в пространстве первых трёх главных компонент. A - лучепёрые рыбы, B - новокрылые насекомые, C – млекопитающие, D – динозавры.

На Рисунке 2 показано распределение геномов митохондрий в пространстве первых трёх главных компонент, при этом на каждом из маленьких рисунков показаны геномы, относящиеся к одной таксономической единице( сверху вниз и слева направо): лучеперые рыбы (A), насекомые (B), млекопитающие (C), архозавры и лепидозавры (D). Хорошо видно, что геномы одной таксономической группы очень плотно располагаются в пространстве собственных компонент( и в естественных координатах — в пространстве частот, соответственно). Можно с очень высокой вероятностью ожидать, что геномы иных организмов, попадающих в таксон соответствующего уровня, также будут плотно располагаться в пространстве частот. При этом справедливо утверждение, «обратное» тому, которое показано на Рисунке 2: если из этого рисунка видно, что геномы одного таксона компактно располагаются в пространстве частот( либо главных компонент), то построение классификации показывает, что один класс полученной классификации преимущественно наполнен геномами одного таксона.

В **первый класс** попадают геномы таксона Actinopterygii : 457 геномов этого таксона из 506 .Во **второй класс** попадают геномы Neoptera ( 137 геномов), 8 геномов Amphibia и единственный геном Mammalia. Сложнее всего устроен **третий класс.** Он содержит 49 геномов таксона Actinopterygii, 39 геномов Amphibia, геномы Archosauria и Lepidosauria – 97 и 87 геномов, соответственно, а также почти все геномы млекопитающих (210 геномов) и Testudines- 24 генома (все внутри одного класса); кроме того, в этот класс попали также 4 генома насекомых.

Анализ распределения геномов по классам показывает, что они распределяются весьма и весьма неслучайно, с точки зрения соответствия класса и таксона. Существенно большая часть одного таксона попадает в один класс, выделяемый по статистическим признакам. Таксон земноводных распределился между двумя классами; примечательно, что по разным классам разошлись в основном бесхвостые (первый класс) и хвостатые земноводные (частично второй и в основном третий класс).
Ключевой задачей для установления связи между структурой генома и таксономией носителя является выявление и анализ триплетов, обеспечивающих разделение на классы.

### 3.2 Исследование симметрии в 32-мерном пространстве

32-мерные подпространства строятся следующим образом: для всех генетических систем показана близость частот слов( цепочек нуклеотидов), составляющих комплиментарные палиндромы: такие пары слов, которые читаются одинаково в противоположных направлениях с учётом замены нуклеотидов по правилу Чаргаффа. Примером такого комплиментарного палиндрома является пара ACTTGC ↔ GCAAGT .
Так из каждой пары, составляющей комплиментарный палиндром, произвольно был выбран один триплет, таким образом, было сформировано два 32- мерных подпространства. Если бы симметрия была идеальной, то построенные классификации ничем бы не отличались. Различия в классификациях- в первую очередь, в составе выделяемых классов- выявляет эффект нарушения симметрии.

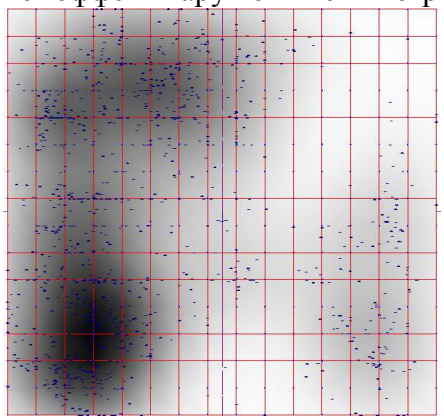

Рис. 5:Распределение 1132 геномов по упругой карте в 32- мерном пространстве частот триплетов, являющихся «левой» половиной комплиментарного палиндрома

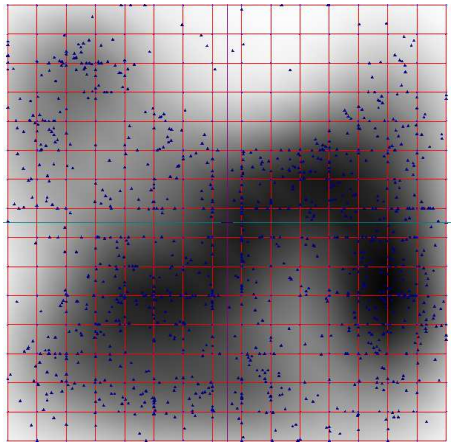

Рис. 6:Распределение 1132 геномов по упругой карте в 32- мерном пространстве частот триплетов, являющихся «правой» половиной комплиментарного палиндрома

Наконец, ещё одно 32-мерное пространство строится следующим образом: это пространство разностей частот

$$r_{\langle \omega,\omega' \rangle} = f_\omega - f_{\omega'} \ ,$$

двух триплетов $\omega$ и $\omega'$, составляющих комплиментарный палиндром. Если бы симметрия была идеальной, то распределение геномов в данном пространстве было бы вырожденным: все координаты всех точек были равны нулю. Нарушение симметрии приводит к тому, что такое вырождение снимается.

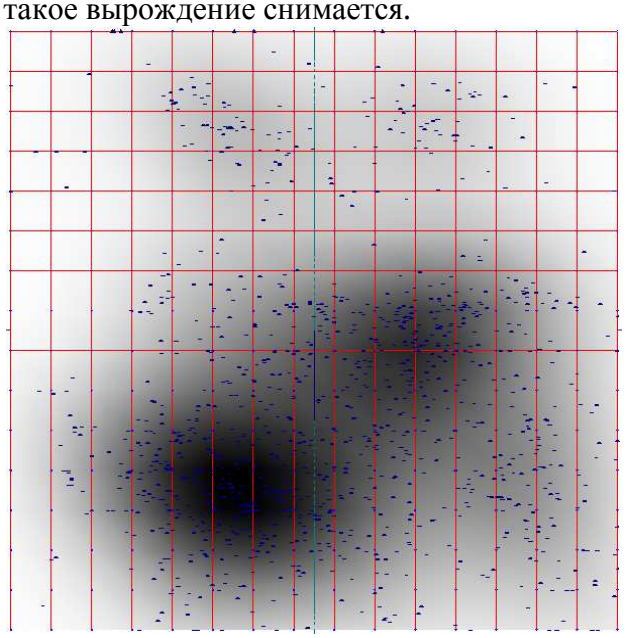

Рис. 4: Распределение 1132 геномов по упругой карте в 32- мерном пространстве разностей триплетов, составляющих комплиментарный палиндром.

Таким образом, можно утверждать, что различные геномы обладают разным уровнем нарушения симметрии Кроме того было выявлено, что не существует чёткой связи между нарушением симметрии и таксономией. Последнее означает, что нарушение симметрии не является продуктом отбора.